SPEAKER VERIFICATION WITH ELICITED SPEAKING-STYLES IN THE VERIVOX PROJECT

I. Karlsson[1], T. Banziger[3], J. Dankovicová[2], T. Johnstone[3], J. Lindberg[1], H. Melin[1], F. Nolan[2], K. Scherer[3]

(1) KTH
Department of Speech, Music and Hearing, Stockholm, Sweden, EU.

(2) CULD
Department of Linguistics, University of Cambridge, Cambridge, United Kingdom, EU

(3) FAPSE
Department of Psychology, University of Geneva, Geneva, Switzerland.

E-mail: verivox@speech.kth.se

ABSTRACT

Some experiments to take care of within speaker variations in speaker verification has been performed. To get speaker variation, speaking behaviour elicitation software has been developed. It was found that if an ASV system was trained on varied speech, speaker verification on even more varied speech improved significantly.

RÉSUMÉ

Nous décrivons les expériences réalisées pour prendre en considération les variations intra-locuteur dans un système de vérification automatique du locuteur (ASV). Afin d'obtenir des variations représentatives de la parole d'un locuteur (différentes vitesses d'élocution, états émotionnels particuliers, etc.), nous avons développé un logiciel spécifique que nous décrivons. Notre travail montre qu'entraîner un ASV avec des échantillons de parole enregistrés dans différentes conditions d'élocution améliore de façon significative les performances du système; et ce, même lorsque le système doit faire face à d'autres types de variations que celles vues lors de l'apprentissage.

1. INTRODUCTION

The aim of the VeriVox project is to improve the reliability of automatic speaker verification (ASV) by developing novel, phonetically-informed methods for coping with the variation in a speaker's voice. Up to now, ASV research has treated within-speaker variation as if it were random, but in fact phonetic research reveals that within-speaker variation is highly structured. Different speaking rates, loudness levels, styles, emotional states, and so on, all cause predictable changes in the acoustic speech signal. The (long-term) goal is to exploit such known phonetic and phonological regularities to reduce the false rejection rate in ASV without a concomitant rise in the risk of false acceptances. This paper reports on the results achieved during the first six month phase of the project, which include a database of various elicited speaking-styles, an acoustic analysis of six of the 50 speakers, and an evaluation with an ASV system.

Our current approach to using phonetic knowledge in an ASV system is called *structured training*. It is the procedure of eliciting different manners of speaking during the enrollment, so that the system becomes familiar with the variation likely to be encountered in that person's voice.

The structured training approach has been tested by comparing it to conventional neutral training using a state of the art HMM-based ASV system developed in the CAVE project [1]. An eliciting software has been implemented and used for collecting a database with 50 male Swedish speakers. The database contains speech to enroll speakers into the system with neutral and with structured training, and further to test the system with a variety of speaking-styles. While recordings are made with a high-quality microphone, full band width and a high sampling rate to allow for various acoustic analyses of the speech, the recordings have been transformed to approximate telephone speech quality for the ASV experiments. This is done because many applications of ASV are expected to appear in the context of telephony, and to prevent research from resulting in methods which are applicable to high-quality recordings only.

This paper is organized as follows. Our implementation of a method to elicit different speech variations to be used in structured training is described in section 2, followed by a summary of preliminary acoustic analyses of recordings from six of the 50 speakers in section 3. The ASV system used in the experiment is described in section 4, while the database and the experiment itself is described in sections 5 and 6. Section 7 presents results from the ASV experiment. We conclude by discussing the results and outlining future research plans.

2. ELICITATION METHOD

The speech database was recorded using a prototype version of eliciting software developed within the project. The software is designed to systematically elicit different types of voluntary and involuntary speech variation. In subsequent trials with an ASV system, the recorded speech samples containing voluntary speech variation are used during the enrollment phase, with the recorded samples containing elicited involuntary speech variation being used as a test set.

Voluntary speech variation is elicited by directing the user to deliberately speak in a number of different modes, including normal, fast, slow, weak, strong and denasalised speech (pinched nose). For each mode, the user is asked to read aloud 6 sequences of 6 digits (2 3 4 5 7 0) in different orders, constructed so that every digit appears in all possible contexts. The order in which the modes are collected is normal, weak, strong, slow, fast, denasalised and normal again.

The software elicits involuntary variation by means of an interactive module in which users perform a succession of tasks which cause them to speak normally, faster and louder without being explicitly asked to do so. The tasks include (i) speaking in the presence of two levels of background white noise (administered through headphones), (ii) speaking from memory at an increased rate due to time pressure and (iii) speaking while solving a divided attention logical reasoning and auditory recognition task, with background noise distraction, allowing the recording of stressed speech. Non-directed normal speech samples are also collected as part of this interactive module. All these tasks are designed to elicit the types of involuntary speech variation which might realistically occur in use of speaker verification systems. This second module (involuntary variation) of the elicitation system uses the same digit sequences as used in the first part (voluntary variation).

3. ACOUSTIC ANALYSIS

Segment durations and formant frequencies at vowel mid-points were measured for 6 speakers saying 3-0-4-5-7-2 (/treː nɔl fyːra fɛm ɧʉː tvoː/) spoken in seven conditions: Neutral, Loud, Weak, Slow, Denasal (pinched nose), and (cognitively) Stressed. One token of each word in each condition was analysed for each speaker, except in the case of Neutral and Stressed where two tokens were analysed. Despite the small amount of data a number of trends emerge, some of which are summarised below.

As expected, all segments in Slow are longer, and most in Fast are shorter. Loud and Weak predominantly involve longer segments. Stressed seems to involve almost consistent shortening of segments.

If each segment's duration is expressed as the percentage change it undergoes as a proportion of the utterance, relative to Neutral, it emerges that a rate change is unevenly distributed over different categories of sound. In Slow there is a clear tendency for the vowels to take up a greater proportion of the lengthening and the consonants less, relative to Neutral; this is also true for Loud. The pattern in Fast is reversed: several consonants take up a greater proportion of the utterance and several vowels take up a smaller proportion.

The first and second formants was also measured. In Fast, with the exception of /fɛm/, the vowels are mid-centralised. In Fast speech there is perhaps less time for the tongue to achieve peripheral articulations. In Slow the vowels are more peripheral. Stressed shows on a smaller scale the (de-peripheralisation) pattern of Fast, while Loud shows in some vowels a pattern of peripheralisation, like Slow. Possibly the changes resulting from Stressed (a difficult style to induce) are similar enough to those resulting from Fast that only Fast need be included in the structured training. Full-scale acoustic results will provide a systematic basis for rationalising the 'structured' training.

The styles tested bring about radical restructuring of the temporal and spectral properties of the speech. This gives a clue to why errors arise in the verification process. Given that 'claim' utterances may differ durationally in complex ways, then even if it is possible to 'time-warp' the claim utterances so that they align well with Neutral reference data, the aligned segments (vowels in particular) will match badly in spectral terms.

4. SPEAKER VERIFICATION SYSTEM

An HMM-based system [1] was used in the experiments. Client models have one left-to-right HMM for each digit. Each HMM has two states per phoneme (there are between two and four phonemes in Swedish digit words) and two Gaussians per state. Speech is parameterised using 12 LPCC coefficients plus an energy coefficient, with appended delta and acceleration coefficients (totally 39 elements per frame). Cepstral mean subtraction is used to decrease inter-session variability. A world model with the same characteristics as the client models is used for log-likelihood normalisation of the score from a client model. An inter-word model (silence and garbage) is shared by all client models and the world model.

When training the world and client models a word boundary segmentation of the training sequences is needed. It is assumed here that an ideal segmentation component is available and this is simulated by using manual segmentations. During the test session the system automatically makes its own segmentations given the sequence of spoken words, i.e., the system knows which words the client actually said.

The system configuration is one of those that performed well in tests in the CAVE project reported on in [1]. The system implementation used in the experiment is described in the same reference.

5. DATABASE

The database used for true-speaker and false-speaker tests in the ASV experiments was collected with the speaking behaviour elicitation software described above. It contains a single 30-minute session from each of 50 speakers which includes both enrollment and verification utterances for the speaker. Given that our interest is mainly in speaker variations due to systematical changes in factors like speaking rate and loudness level, we found it reasonable for a first study to use material from a single recording session.

All speakers in the database are male and come from the same (broad) dialect region around Stockholm. The speech material used in the ASV experiments consists of sequences of six digits spoken in Swedish. Each such sequence contains the digits 0, 2, 3, 4, 5 and 7 in various orders.

Recordings were made in a sound-treated booth with a high-quality head set microphone and a sample rate of 22 kHz. These full bandwidth recordings are used for various acoustic analyses within the project. For a first order approximation of telephone speech quality in the ASV experiments, recordings were down-sampled to 8 kHz, band-pass filtered to approximately telephone bandwidth (300-3400 Hz), and finally quantized to 8-bit A-law coding (ITU G.711). Digit sequence boundaries were then marked manually, as well as word boundaries for the enrollment speech.

For training the world and silence models in the ASV system, recordings from 15 male and 15 female speakers in a separate telephone speech database [2] were used. In this way the

ASV system is setup to work with telephone quality speech and with impostors of both genders.

6. EXPERIMENT

The purpose of the experiment is to test the hypothesis that structured training, as defined in the introduction, is helpful in making an ASV system more robust to naturally occurring variations in speaking-style, and especially to reduce the false rejection rate at a given false acceptance rate. Two enrollment sets were therefore defined from the first part of the session, set A and set B. Set A represents conventional *neutral training* and contains only neutral speech. It serves as a baseline in the experiment. Set B simulates *structured training* and contains equal amounts of all six speaking-styles included in the first part of the session (Neutral, Weak, Strong, Slow, Fast and Denasal). Both enrollment sets have equal size and contain 12 six-digit sequences each.

To compare the two enrollment sets, one batch of tests was run for each set. For each enrollment set, all 50 speakers were first enrolled as clients in the system, and a set of 31 true-speaker and 49 false-speaker tests per client was then performed. The set of true and false speaker tests is identical for both batches.

Verification utterances for the simulated identity claims were taken from the second part of each speaker's session, and each verification test was made with one six-digit utterance. For true-speaker claims, all utterances from the second part were used, and Table 1 shows the distribution of these test utterances over speaking-style. The full set of true-speaker test utterances is called the *Composite* set and contains a somewhat realistic, but pessimistic, mix of speaking-styles that could appear during use of an ASV system. When analyzing results from the experiments, three disjoint subsets of the Composite set will also be referred to, whose composition is also shown in Table 1.

For simulated impostor attempts, one neutral speech utterance from each speaker was selected for an attempt against all other speakers' identities, yielding 49 independent impostor attempts per enrolled client. The exact same series of impostor attempts were used with each of the partitions of the set of true-speaker tests.

7. RESULTS

The main objective of using structured training is to reduce false rejection rate for a given false acceptance rate. Such reductions for various operating points can be read from a detection error trade-off (DET) curve [3], and Figure 1 shows such curves for the two enrollment sets and the Composite test set. The reduction in false rejection rate for structured training compared to neutral training, at a fixed false acceptance rate, is the vertical distance between the two corresponding DET curves. If for example, we start with the equal-error-rate point at neutral training, the false rejection rate is reduced from 2.7% to 1.4% when changing to structured training: a 48% reduction in error rate.

Figure 2 shows DET curves for each of the different partitions of the test set with neutral training (2a) and with structured training (2b). Note that many of the speaking-

Style	*Composite*	*Neutral*	*Stress*	*Other*
Neutral	16	16		
Weak	1			1
Strong	1			1
Slow	1			1
Fast	1			1
Denasal	1			1
Noise, weak	1			1
Noise, loud	2			2
Memory, fast	1			1
Stressed	6		6	
total size:	31	16	6	9

Table 1. The number of true-speaker tests of each elicited speaking-style per client. The *Neutral*, *Stress* and *Other* sets are disjoint subsets of *Composite*.

Enrollment	*Composite*	*Neutral*	*Stressed*	*Other*
A/neutral	2.66	1.26	1.67	4.77
B/structured	1.80	1.26	2.29	2.46

Table 2. Average (same-sex) equal-error-rate for the two enrollment sets over each of the test sets. Thresholds are speaker-independent and calculated a posteriori.

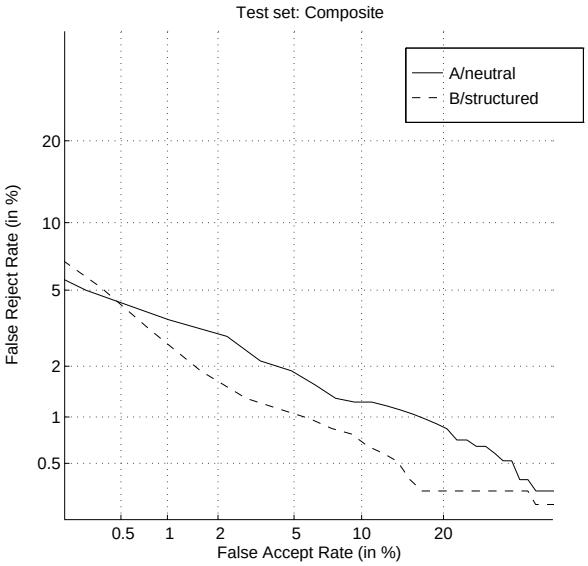


Figure 1. DET curves for the system with conventional neutral training and with structured training, tested on the Composite test set. The threshold parameter is speaker-independent.

styles included in the *Other*-partition are also included in the structured training and that there is a corresponding improvement. For the *Stress*-partition, on the other hand, performance is slightly worse with structured training than with neutral training, and it seems that structured training has not captured variations introduced with elicited stress.

Table 2 finally, summarizes the results in average equal-error-rates with speaker independent thresholds for each of the two enrollment sets and the different partitions of the test set.

8. DISCUSSION

With neutral training the error rates for the different test sets seem to move apart more than for the same tests when using structured training. This means that the neutrally trained models give a varying performance for the different elicited speaking-styles, while the structured training causes a more similar performance for the different speaking-styles. This without deteriorating the performance for the Neutral test set.

With a world model built from neutral training and with structured training of the client model, it is likely that the system would be poor in rejecting impostor attempts with non-neutral speech. The client model is trained to match a broader variation of speech than with neutral training (as long as it is not very well tuned to the particularities of the modeled speaker), while the world model is a poor model of impostors speaking in a non-neutral manner. It is therefore important that in a system with structured training, the world (or cohort) model is also created with structured training. In the current experiment, this problem was circumvented by using impostor attempts with neutral speech only.

9. CONCLUSION

In the first phase of the VeriVox project, structured training has been tested as a way of making a speaker model in the ASV system familiar with variations in a speaker's voice likely to be encountered in future access attempts. A near halving of the average false rejection rate was demonstrated on a mixed speaking-style test set, at no increase in false acceptance rate, and this clearly shows the feasibility of the approach.

So far the original ASV system itself has not been modified. In phase two of the project we will look for ways to improve the system's robustness to variations in speaker style, for instance by modifying the speaker model to better make use of the data seen through structured training. Another possibility is "guided elicitation", whereby the system guides a client into the right way of speaking in case of a negative outcome from a first verification test.

10. ACKNOWLEDGEMENT

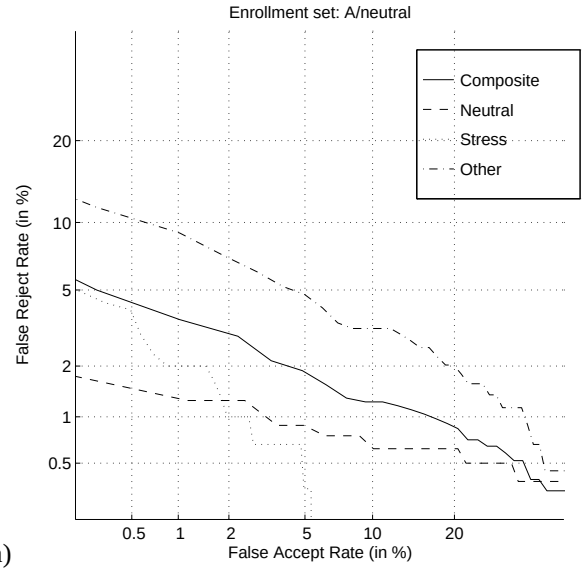
The first phase of VeriVox was supported by the EU Esprit program. Partners of the consortium are Cambridge Univ. (UK), CNRS (F), KTH (S), Queen Margaret College (UK), Univ. Of Dublin (IRL), Univ. Bonn (D), Univ. Geneva (CH), and Ensigma Ltd. (UK). Special thanks to Lennart Nord (KTH) for assistance with the database recording.

REFERENCES

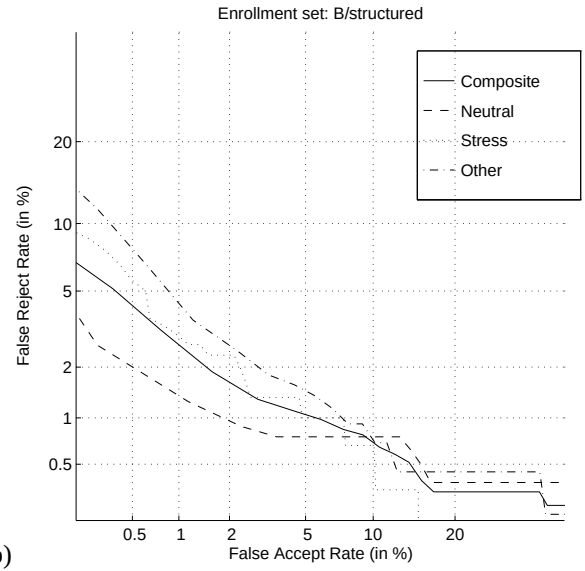
[1] Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J., and Pierrot J.-B., "Speaker Verification in the Telephone Network: Research activities in the CAVE Project," *Proc. EUROSPEECH'97*, Rhodes, Greece, 1997, Vol. 2, pp971-974.

[2] Melin H., "Gandalf - A Swedish Telephone Speaker Verification Database," *Proc. ICSLP-96*, pp. 1954-1957, Philadelphia, USA, 1996.

[3] Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M., "The DET Curve in Assessment of Detection Task Performance", *Proc. EUROSPEECH'97*, Rhodes, Greece, 1997, Vol. 4, pp. 1895-1898.



a)



b)

Figure 2. DET curves for the Composite test set and the three subsets of it, with a) neutral training, b) structured training. The threshold parameter is speaker-independent.